JSLHR, Volume 40, 900–911, August 1997

Relationships Between Subjective Ratings and Objective Measures of Performance in Speechreading Sentences

Marilyn E. Demorest
University of Maryland Baltimore County

Lynne E. Bernstein*
Center for Auditory and Speech Sciences
Gallaudet University
Washington, DC

Ninety-six participants with normal hearing and 63 with severe-to-profound hearing impairment viewed 100 CID Sentences (Davis & Silverman, 1970) and 100 B-E Sentences (Bernstein & Eberhardt, 1986b). Objective measures included words correct, phonemes correct, and visual-phonetic distance between the stimulus and response. Subjective ratings were made on a 7-point confidence scale. Magnitude of validity coefficients ranged from .34 to .76 across materials, measures, and groups. Participants with hearing impairment had higher levels of objective performance, higher subjective ratings, and higher validity coefficients, although there were large individual differences. Regression analyses revealed that subjective ratings are predictable from stimulus length, response length, and objective performance. The ability of speechreaders to make valid performance evaluations was interpreted in terms of contemporary word recognition models.

KEY WORDS: speechreading (lipreading), word recognition, subjective ratings

The ability to monitor one's reception of connected discourse and to evaluate whether a message has been correctly understood is one component of effective communication. When communication takes place under favorable acoustic and optical conditions, between individuals with normal hearing and vision, metalinguistic processes such as comprehension monitoring, although likely ongoing, play a less critical role than under less favorable conditions. If the acoustic speech signal is degraded because the listener has a hearing impairment, the ability to detect and correct communication breakdowns takes on increased importance. This is also true when communication is based primarily on the optical speech signal, as in speechreading. The present investigation examined the validity of subjective ratings of performance in a speechreading task for individuals with normal and impaired hearing. Also, evidence was sought to explain how perceivers, in the absence of external feedback, are able to produce valid subjective ratings. The obtained evidence points to several different information sources, including, possibly, access to activation during word recognition.

Self-generated performance ratings have been investigated in domains as diverse as psychophysics and educational achievement (for

*Lynne E. Bernstein is currently affiliated with the House Ear Institute, Los Angeles, California.

 1092-4388/97/4004-0900

examples, see Green & Swets, 1966, and Wang & Stanley, 1970), and findings converge on the conclusion that judgments are generally accurate. That is, subjective performance ratings are monotonically related to objective measures of performance. Consistent with these findings, in the literature on auditory speech recognition testing, it has been demonstrated repeatedly that participants can estimate the percentage of running speech they can understand and can adjust speech or competing background noise to a specified level of intelligibility. In the first such study, Speaks, Parker, Harris, and Kuhl (1972) found that ratings of auditory intelligibility[1] of CID Everyday Sentences (Davis & Silverman, 1970) by participants with normal hearing were highly correlated with objective measures of percent words correct. Subsequently, Gray and Speaks (1978) showed that listeners with hearing impairment could also reliably rate speech intelligibility. The cumulative evidence leads to the expectation that participants should also be able to provide valid estimates of their performance in a speechreading task.

For those with hearing impairment who rely heavily on speechreading for communication, performance monitoring is likely to be a more important part of everyday communication than for those with normal hearing, because opportunities for miscommunication abound. In addition, it is quite likely that individuals in the former group will have had explicit training in speechreading, with feedback, and that they will therefore be more experienced evaluators of their own performance. For these reasons, it could be anticipated that the correlation of subjective ratings and objective measures of speechreading performance would be higher for individuals with hearing impairment. This prediction contrasts with the data from a small, but relevant, study on auditory speech perception. Yanz and his colleagues (Yanz, 1984; Yanz, Carlstrom, & Thibodeau, 1985) used a measure of self-assessment ability derived from signal detection theory to compare groups of participants with normal and impaired hearing in an auditory word-identification task. The group with hearing impairment identified fewer words correctly and also demonstrated slightly lower self-assessment accuracy, but the latter difference was not significant. Given the small sample sizes (n = 10 in each group) and consequent lack of statistical power, the finding is not conclusive, but the difference between the sample means was the opposite of what we would predict for speechreading.

It has been shown that there are large individual differences in speechreading ability, regardless of hearing status (Bernstein, Demorest, & Tucker, 1996; Demorest & Bernstein, 1992; Demorest, Bernstein, & DeHaven, 1996; Demorest, Bernstein, & Tucker, 1997). Bernstein et al. (1997) have argued that the high level of performance of some individuals with hearing impairment demonstrates that speech perception can be acquired in the absence of audition. The factors that account for such large differences have yet to be identified. We assume that factors may be found at several levels, including perceptual, linguistic, and cognitive. If subjective ratings are a type of, or source for, internal feedback, better speechreading may be associated with a more reliable internal feedback mechanism. It was therefore of interest in the present study to examine individual differences in the ability to make valid ratings of performance.

If subjective estimates are correlated with objective performance measures, a question that arises is how participants are able to provide such estimates in the absence of objective feedback. That is, how do participants judge their own accuracy in the absence of additional non-perceptual information? Regression analysis was used to determine what aspects of the stimulus and of the participant's response are related to the subjective ratings. The results provide a statistical model of the performance ratings that has theoretical implications, which are raised in the Discussion section.

In summary, the present investigation examined: (a) the validity of subjective ratings of speechreading performance in terms of objective performance; (b) group and individual differences in the validity of the ratings; and (c) the stimulus and response measures that are predictive of the ratings, and possibly indicative of the bases for subjective ratings.

A general theme in the speechreading literature is the requirement to account for performance on both perceptual and cognitive/linguistic grounds. Our strategy has been—as is true here—to examine large samples of performance employing descriptive techniques as a precursor to examining perceptual and cognitive processes. Our belief is that it is a prerequisite to find out what people can do and then address what accounts for their performance.

Method

Participants

Participants were 96 adults with normal hearing who were graduate or undergraduate students at the University of Maryland Baltimore County and 63 adults with severe-to-profound hearing impairment who were undergraduate students at Gallaudet University. Descriptive statistics and psychometric analyses of these individuals' speechreading performance on CID Sentences (Davis &

[1]The term *speech intelligibility* is usually used when the characteristics of a talker or speech communication system are being evaluated, but it has also been used when a participant's identification accuracy is being measured (Duffy & Pisoni, 1992) Present usage of the term conforms to the literature being cited.

Silverman, 1970), Bernstein-Eberhardt (B-E) Sentences (Bernstein & Eberhardt, 1986b), words (Modified Rhyme Test, Kreul et al., 1968), and nonsense syllables have been reported by Demorest et al. (1996), Bernstein et al. (1997), and Demorest et al. (1997). Although 72 individuals participated in the last study, sample size for the present analyses is smaller (n = 63) because the rating task was not introduced until after several participants had been tested.

Participants with normal hearing ranged in age from 18 to 45 years (M = 22.6); those with impaired hearing ranged from 18 to 41 years (M = 23.3). Additional selection criteria for the latter group included having families with English as the native language, education in a mainstream and/or oral setting for a minimum of 8 years, and no additional impairments. All participants had normal or corrected vision and English as a native language. A majority of the participants with impaired hearing (69.5%) had profound hearing losses of early onset (before 36 months).

Materials

Stimuli were video laserdisc recordings of 100 CID Everyday Sentences (Bernstein & Eberhardt, 1986a) and 100 sentences from Corpus III and IV of Bernstein and Eberhardt (1986b) (B-E Sentences). Fifty sentences in each set were spoken by a male talker and 50 by a female talker. Details of the recording procedures can be found in Demorest and Bernstein (1992).

Participants with normal hearing viewed all 200 sentences in two 2-hr sessions, 5 to 12 days apart. On Day 1, 50 sentences from each set (25 from each talker) were presented. The remaining sentences were presented on Day 2. Order of talkers and sentence sets was counterbalanced across participants. Participants with hearing impairment were tested in a single session during which 100 sentences (25 from each set and talker) were presented.

Procedure

General testing procedures are described in detail in Demorest et al. (1996). Briefly, participants were seated in a darkened, sound-attenuated room. A laboratory computer presented the stimuli via a laserdisc player (Sony Lasermax LDP 1550) and a high-quality video monitor (Sony Trinitron) and collected the participant's typed response via the computer keyboard. Participants were informed that they would see a series of unrelated sentences and were instructed to type exactly what they thought the talker said. Partial responses, including word fragments, were encouraged, but participants were instructed to keep all portions of their response in chronological order.

The most common approach to subjective performance evaluation in psychophysical and cognitive tasks has been to define subjective performance operationally as a confidence rating. Confidence ratings have the advantage that they are readily understood by participants and can be obtained efficiently. Accordingly, following the response to each sentence, the participant was instructed to "rate your confidence in the correctness of your response." Anchors for the rating scale were: 0 = *no confidence—I guessed* and 7 = *complete confidence—I understood every word*. Numbers between 0 and 7 were used to represent intermediate degrees of subjective performance. The zero rating included the phrase "I guessed" so as to acknowledge that participants had been encouraged to respond and that they might therefore have lowered their criteria for responding (Van Tasell & Hawkins, 1981). Five practice sentences were given before data collection began, at which time questions or misunderstandings concerning the procedures or rating scale were addressed.

Data Preparation

Participants' typed responses were reviewed for obvious typographical errors and corrected, provided there was no ambiguity whatsoever concerning the intended response (e.g., "feeleing" for "feeling" or "allready" for "already"). Responses were also transcribed into a quasi-phonemic notation using the DECtalk text-to-speech synthesizer (DECtalk DTC01, Version 2.0; Educational Services Department, Digital Equipment Corporation, 1984). See Bernstein, Demorest, and Eberhardt (1994) for additional details.

Measures

Three measures were based on word-level scoring: (a) the number of *stimulus words*; (b) the number of *response words*; and (c) the number of *words correct*. These measures were extracted using software that operated on the edited text of the stimulus and of the participant's response.

Several measures were based on a phonemic analysis of the responses. Determination of the number of *phonemes correct* required a strategy for aligning the phonemes of the stimulus sentence with the phonemes of a participant's response. Although this is impossible to do manually when there are a great many errors, sequence comparison algorithms (Kruskal & Sankoff, 1983) implemented in software were used to generate such stimulus-response alignments. A sequence comparator described by Bernstein et al. (1994) was developed for this purpose, and its application to speechreading data was illustrated by Demorest and Bernstein (1991). Briefly, the comparator examined all possible

alignments of a stimulus-response pair and selected the alignment that minimized overall visual phonetic distance between the stimulus and response. Visual distance was calculated on a phoneme-to-phoneme basis and summed across all elements of the alignment. Visual-phonetic distances used by the comparator were based on multidimensional scaling of nonsense-syllable confusions, theoretical considerations, and criteria for the performance of the comparator (see Bernstein et al., 1994).

As an example, consider the stimulus sentence "Why should I get up so early in the morning?" and the response "Watch what I'm doing in the morning!" The sequence comparator aligned the phonemes of these two sentences as follows:

```
Stimulus:  wASUdAgEt^psoRliInDxmorn|G
Response:  waC---wxtAmdu|G-InDxmorn|G
```

Inspection of the alignment reveals that there were 26 *stimulus phonemes* and 22 *response phonemes*. Twelve of the response phonemes were "correct," that is, aligned with the same phoneme in the stimulus. Also included as a performance measure was the normalized (or average) visual distance between the stimulus and response. *Normalized visual distance* was calculated by adding the visual distance associated with each position in the alignment and dividing by the number of stimulus phonemes. The metric for distances between individual phonemes ranged from 0 to 42. In the above example, normalized visual distance = 14.1. Non-responses were scored as 0 phonemes correct. The sequence comparator assigned the value 8.0 to insertions or deletions (see Bernstein et al., 1994, for a complete discussion of the comparator). Thus, non-responses received a visual distance of 8.0, which was equivalent to deletion of all stimulus phonemes. Visual distance measures are expected to correlate negatively with performance ratings and with other objective measures of performance because phonemes that are correct have values of 0 for visual distance, and large visual-phonetic distances represent poor performance.

Results

Descriptive Statistics

Means and two measures of variability for all stimulus and response measures are shown in Table 1, for each set of sentences and for each group. Individual differences among participants are reflected in the between-subjects standard deviations (SD_B): For each measure, a mean was calculated for each participant, across sentences, and then the standard deviation of the participant means was calculated. Variability within participants, across sentences, is described by within-subjects standard deviations (SD_W): For each measure, a variance was calculated across sentences for each participant, the variances were averaged across participants, and the square root was then taken.

Means

As Bernstein et al. (1997) have reported, mean performance on the speechreading task was higher in the group of participants with hearing impairment. There was also a corresponding difference in confidence: $t(157) = 6.91$, $p < .0005$, for CID Sentences; $t(157) = 7.67$, $p <$

Table 1. Means and standard deviations of stimulus and response measures as a function of sentence set and group.

	CID sentences						B-E sentences					
	Normal-hearing			Hearing-impaired			Normal-hearing			Hearing-impaired		
Measure	*M*	SD_B	SD_W	*M*	SD_B	SD_W	*M*	SD_B	SD_W	*M*	SD_B	SD_W
Confidence	2.64	1.14	2.02	3.99	1.29	1.72	2.07	1.02	2.17	3.47	1.27	2.16
Stim. wd.	7.49	0.00	3.20	7.62	0.00	1.91	6.04	0.00	3.06	5.76	0.00	1.96
Stim. ph.	22.98	0.00	11.30	22.91	0.00	7.38	20.21	0.00	10.57	19.57	0.00	7.62
Resp. wd.	4.16	1.39	2.76	5.35	1.45	2.23	3.69	1.27	3.07	4.16	1.06	2.16
Resp. ph.	11.13	3.83	7.68	15.22	4.14	6.37	10.19	3.50	9.00	12.77	3.28	6.81
Wd. cor.	1.86	0.96	2.14	3.23	1.49	1.47	1.20	0.51	2.68	2.22	0.88	1.84
WCNS	.28	.13	.32	.45	.19	.32	.20	.08	.25	.40	.15	.32
Ph. cor.	6.16	2.82	5.94	10.74	4.27	4.50	4.67	1.78	8.00	8.38	2.93	6.10
PCNS	.33	.13	.30	.52	.18	.31	.26	.10	.25	.47	.16	.31
VDNS	6.61	1.37	3.12	4.84	1.66	2.49	7.38	1.06	3.07	5.34	1.47	3.04

Note. Stim. wd. = stimulus words; Stim. ph. = stimulus phonemes; Resp. wd. = response words; Resp. ph. = response phonemes; Wd. cor. = words correct; WCNS = normalized words correct; Ph. cor. = phonemes correct; PCNS = normalized phonemes correct; VDNS = normalized visual distance. SD_B = between-subjects standard deviation; SD_W = within-subjects standard deviation.

.0005, for B-E Sentences. Although the performance ratings were made on an arbitrary 0 to 7 scale, the anchor points were defined in such a way that conversion of the ratings to percentages provides a *rough* estimate of the subjectively experienced level of performance. The four mean confidence ratings in Table 1 convert to 37.7%, 57.0%, 29.6%, and 49.6%, respectively. These values are slightly higher than, but remarkably similar to, the mean percentage of phonemes correct per sentence (see means for normalized phonemes correct in Table 1).

Standard Deviations

Between-subjects standard deviations are zero for the stimulus measures because all participants received the same sentences. Within-subjects standard deviations are non-zero, but are the same for every participant. In generalizability analyses of these data for the participants with normal hearing, Demorest, Bernstein, and DeHaven (1996) have shown that the participant and the sentence are both significant sources of variability in performance on individual sentences. The smaller magnitude of the between-subjects standard deviations reflects, in part, the fact that they are based on mean performance across either 100 or 50 sentences.

Validity: Analysis of Non-Responses

If participants correctly interpreted the rating scale, non-responses should have received ratings of 0. Therefore, as a preliminary check on the validity of the ratings, non-responses were examined. Across the 25,500 sentences presented (200 sentences for each of 96 participants; 100 for each of 63 participants), there were 3,574 trials on which participants typed no response, but simply provided a confidence rating. Of these, 3,557 (99.5%) were rated 0.

Validity: Within-Subjects Correlations

The overall validity of the confidence ratings was evaluated separately for the two groups of participants. Within each group, ratings of individual sentences were nested within subjects. To obtain a measure of association between the ratings and performance that was independent of differences in level of performance from one individual to the next, a pooled, within-subjects correlation was obtained between the ratings and each objective measure.[2] Calculation of within-subjects correlations is based on sums of squares and sums of cross-products that are taken around each participant's mean and summed (pooled) across participants.[3] Conceptually, within-subjects correlations represent the relations between the variables after partialing out the individual differences in participant means. These correlations provide estimates of the validity of the participants' ratings within each group and will henceforth be termed *validity coefficients*.

Validity coefficients were calculated between the subjective ratings and five objective performance measures: words correct (WC), words correct normalized on stimulus length (WCNS), phonemes correct (PC), phonemes correct normalized on stimulus length (PCNS), and visual distance normalized on stimulus length (VDNS). Because the zero ratings given to non-responses served to anchor the confidence rating scale, all sentences were included in these correlations. For participants with normal hearing, separate coefficients were calculated for the 100 CID Sentences and the 100 B-E Sentences. For participants with hearing impairment, separate coefficients were calculated for the 50 sentences in each set.

The validity coefficients, which are presented in Table 2, support the prediction that participants with and without hearing impairment can provide valid ratings of their performance on a speechreading task. The magnitude of the correlations is impressive for both groups, but especially so for the participants with normal hearing, most of whom reported never having attempted to speechread before. Magnitude of the correlations is consistently higher for the normalized measures of words and phonemes correct than for the unnormalized measures. An interpretation of this difference is that participants' ratings represent a subjective evaluation of proportion (or percentage) of the sentence that was correct. This interpretation is consistent with the instructions given to the participants and with the finding above that the mean ratings can be mapped onto a percentage scale that agrees closely with the percentage of phonemes correct per sentence.

Group and Individual Differences in the Validity of Performance Ratings

The correlations in Table 2 are consistently higher for the participants with hearing impairment, as predicted. That is, they were better able to judge their

[2]Finn (1974, pp. 81–83) and Harris (1985, pp. 180 & 250) discuss the impact of performing correlational analyses on total versus within-groups statistics when there is a subgroup structure to a set of data.

[3]The definitional formula for the within-subjects correlation between variables X and Y is given by

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{K}(X_{ij}-\overline{X}_i)(Y_{ij}-\overline{Y}_i)}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{K}(X_{ij}-\overline{X}_i)^2\sum_{i=1}^{N}\sum_{j=1}^{K}(Y_{ij}-\overline{Y}_i)^2}},$$

where X_{ij} and Y_{ij} are measures for subject i on sentence j, N is the number of subjects, and K is the number of sentences.

Table 2. Validity coefficients for subjective ratings and objective measures of speechreading performance as a function of sentence set and group.

Measure	CID Sentences		B-E Sentences	
	Normal-hearing	Hearing-impaired	Normal-hearing	Hearing-impaired
Words correct	.506	.521	.533	.616
Normalized words correct	.676	.732	.565	.694
Phonemes correct	.476	.483	.561	.628
Normalized phonemes correct	.711	.764	.616	.732
Normalized visual distance	−.511	−.626	−.339	−.573

Note. All correlations are statistically significant, $p < .0001$.

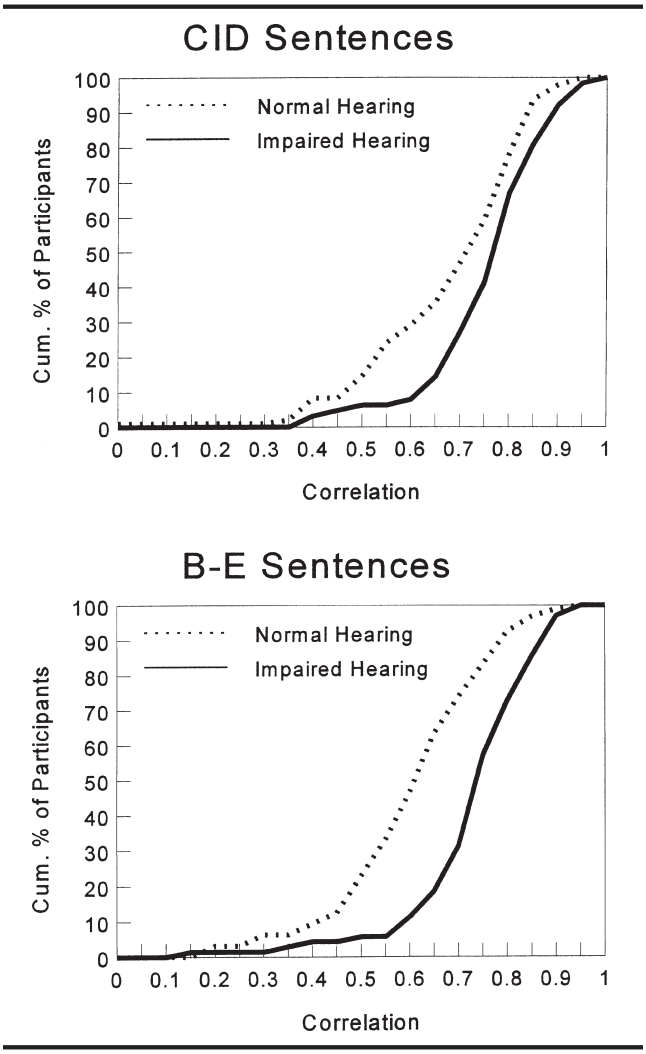
performance on the speechreading task. Fisher's Z transformation (Steel & Torrie, 1980, p. 279) was used to compare the correlations for the two groups. With a Bonferroni correction for multiple statistical tests, eight of the 10 comparisons were significant ($p < .005$). The exceptions were the correlations for words correct and phonemes correct in CID sentences.

Because the coefficients presented in Table 2 were obtained by pooling sums of squares and cross-products across participants, they do not provide information about individual differences in the validity of the performance ratings. To examine individual differences, within-subject correlations (i.e., validity coefficients) were calculated, for each participant, between the performance ratings and the five objective measures of performance. Median correlations and ranges are presented in Table 3 for each sentence set and group.

The medians of the distributions of validity coefficients are very similar to the overall validity coefficients shown in Table 2, and either statistic could be used as a summary statistic. For participants with normal hearing, a coefficient greater than .196 in magnitude is statistically significant (100 sentences, $df = 98$, $p < .05$). For the participants with hearing impairment, the critical value is .278 (50 sentences, $df = 48$). Across the five measures, for the group with normal hearing, 93.0% of the coefficients were statistically significant in the expected direction. For participants with impaired hearing, 96.4% of the coefficients were significant. Thus, in both groups, nearly all participants were able to provide valid (i.e., non-chance) ratings of their performance.

To illustrate further both group and individual results, cumulative frequency distributions (%) for each sentence set are shown in Figure 1 for the normalized measure of phonemes correct (PCNS). This measure was chosen because the mean performance ratings were most similar to PCNS and because the validity coefficients for PCNS were the highest. The percentage values can be interpreted as the percentile equivalents of the correlations. That is, the figures show the percentage of participants with a correlation less than, or equal to, a given value. The distributions clearly show the

Figure 1. Cumulative frequency distributions of within-subjects correlations between confidence and normalized phonemes correct for participants with normal hearing ($N = 96$) and participants with impaired hearing ($N = 63$). Top panel shows results for CID Sentences; bottom panel shows results for B-E Sentences.



superiority of the participants with impaired hearing, more than 90% of whom have coefficients ≥ .600. Low coefficients are more frequent among the participants with normal hearing, but 90% of their coefficients exceed .400. Thus, for both groups, the validity coefficients are not only statistically significant, but are also quite strong.

Despite the moderately high validity coefficients overall, there is an extremely wide range across participants. Coefficients for some participants in each group exceed .900, whereas for others, the coefficients are near zero, and some even have an inappropriate sign (see Figure 1 and the ranges in Table 3). Although it would be interesting to know whether the validity of subjects' performance ratings is related to their overall level of performance, the present data cannot be used to test this hypothesis because of range artifacts. Participants with extremely poor performance tended to have low variability in both performance and in their confidence ratings. Thus, their validity coefficients tended to be affected by restriction of range in both measures. A better way to determine whether speechreading ability is related to the ability to evaluate one's performance would be to adjust the difficulty of the speechreading material for individual participants while assessing the validity of the ratings and to hold materials constant while assessing speechreading ability.

Within-Subjects Regression Analyses

The results presented thus far demonstrate convincingly that both experienced and inexperienced speechreaders are able to judge their own performance on these sentence materials. That they can do so, in the absence of feedback, suggests that relevant information is available to them without specific training. Two sources of information that participants might use in making their confidence ratings are the length of the stimulus and the length of the response. Stimulus length may be a determiner of subjectively evaluated difficulty of the sentence. On the average, participants may consider the sentences less difficult and have greater confidence when the stimulus is short (e.g., "Good morning," CID Sentence 5) than when it is long (e.g., "The morning paper didn't say anything about rain this afternoon or tonight," CID Sentence 69). Such an assumption has validity: Across sentence sets and groups, pooled within-subjects correlations of stimulus length with WCNS and PCNS ranged from -.084 to -.324. The longer the sentence, the smaller the proportion of words and phonemes correct.

It is also plausible that the length of the participant's response would be related to the subjective ratings. First, as more and more words are given in a response, the a priori probability of correct responses increases: Pooled within-subjects correlations of response length with WC and PC (the total number of words and phonemes correct) range from .549 to .851. Second, it is possible that participants have a decision criterion for responding (Van Tasell & Hawkins, 1981) and that the number of words or phonemes in the response reflects the number that have exceeded that criterion. Although this hypothesis cannot be tested directly in the present study, it is consistent with the observation that nonresponses nearly always receive a zero confidence rating. Thus, even without feedback on the correctness of a given response, other things being equal, a participant can estimate that a

Table 3. Medians and ranges of participants' validity coefficients for objective measures and subjective ratings of speechreading performance as a function of sentence set and group.

	CID Sentences		B-E Sentences	
Measure	Normal-hearing	Hearing-impaired	Normal-hearing	Hearing-impaired
		Median		
Words correct	.505	.531	.535	.640
Normalized words correct	.690	.742	.560	.710
Phonemes correct	.468	.496	.565	.650
Normalized phonemes correct	.730	.777	.615	.720
Normalized visual distance	−.512	−.674	−.340	−.610
		Range		
Words correct	−.083 to .794	.202 to .801	.078 to .876	.006 to .837
Normalized words correct	−.036 to .929	.354 to .917	.105 to .903	.011 to .893
Phonemes correct	−.120 to .789	.089 to .744	.071 to .855	.081 to .916
Normalized phonemes correct	−.031 to .935	.373 to .958	.169 to .912	.129 to .931
Normalized visual distance	−.877 to .169	−.900 to −.105	−.807 to .215	−.908 to .002

Note. Ranges are reversed for normalized visual distance because it correlates negatively with performance.

longer response is more likely to result in a larger proportion of words or phonemes correct. Hence it is reasonable to expect that response length would be predictive of the participant's confidence rating.

To understand better the possible bases for participants' performance ratings and to determine whether objectively measured performance still correlates with the ratings after controlling statistically for stimulus and response length, the ratings were used as dependent variables in a series of multiple regression analyses. Several approaches to defining the predictor variables were examined. Although normalized measures of words and phonemes correct had higher zero-order correlations with subjective ratings than unnormalized measures, using more than one normalized measure in the same analysis results in a high degree of collinearity among the predictors. This is undesirable because it makes the regression coefficients unreliable and therefore makes it difficult to evaluate the unique contribution of each predictor. For this reason, unnormalized measures of correct performance were used.

Another source of collinearity arises from the high correlation between words correct and phonemes correct. Although, strictly speaking, this is not a part-whole correlation, it behaves similarly because each correct word necessarily increases the number of phonemes correct. To avoid this problem, separate analyses were performed with word and phoneme measures. An additional advantage of this separation is that the word measures can be obtained with conventional scoring, whereas the phoneme-based measures require use of the sequence comparator. Results from the word-based analyses would therefore have broader usefulness than those from the phoneme-based analyses.

Two regression analyses were performed for each sentence set in each group. In the first, the number of stimulus words, response words, and words correct were the predictors. In the second, the number of stimulus phonemes, response phonemes, and phonemes correct were predictors, together with the normalized measure of visual distance. Although the number of stimulus phonemes appears in the denominator of the latter measure, it is important for conceptual reasons to normalize visual distance so that it represents the degree of visual dissimilarity between the stimulus and response independent of sentence length.

Regression analyses were based on the within-subjects correlation matrix. As noted previously, these correlations are obtained by combining within-subject sums of products and sums of squares. The degrees of freedom for a within-subjects correlation equal the number of observations (number of sentences x number of participants) minus the number of participants.

Results of the regression analyses are summarized in Table 4. For each analysis, the zero-order correlation with confidence and the standardized regression coefficient (β), if significant at $p < .05$, are given for each predictor. Also shown is the obtained multiple correlation, R, and its square, which represents the proportion of sample variance explained. Across the eight analyses, values of R ranged from .637 to .776, and variance accounted for ranged from 40.6% to 60.3%. With one exception, all regression coefficients were statistically

Table 4. Zero-order within-subjects correlations and regressions of performance ratings on stimulus and response measures as a function of sentence set and group.

	CID Sentences				B-E Sentences			
	Normal-hearing		Hearing-impaired		Normal-hearing		Hearing-impaired	
Measure	*r*	β	*r*	β	*r*	β	*r*	β
Word measures								
Stimulus words	−.288	−.453	−.199	−.610	−.101	−.218	−.165	−.346
Response words	.369	.220	.390	.342	.523	.355	.539	.304
Words correct	.506	.453	.521	.520	.533	.372	.616	.492
R		.664		.725		.637		.724
R^2		.441		.526		.406		.524
Phoneme measures								
Stimulus phonemes	−.309	−.425	−.205	−.468	−.151	−.268	−.147	−.326
Response phonemes	.347	.388	.361	.560	.527	.391	.558	.480
Phonemes correct	.476	.078	.483		.561	.239	.628	.095
Normalized visual distance	−.511	−.361	−.626	−.461	−.339	−.130	−.573	−.337
R		.708		.766		.652		.754
R^2		.501		.587		.425		.569

Note. All correlation and regression coefficients are statistically significant, $p < .0001$.

significant. In the sample with hearing impairment, for phonemes correct, β = .072, p = .058. Zero-order correlations and regression coefficients for the measures of stimulus length are all negative, indicating that ratings are lower, as expected, when the stimulus sentence is longer. Measures of response length correlate positively with ratings, indicating that when participants are confident they produce more response words and phonemes.

Perhaps the most important aspect of the regression results is that objectively measured performance continues to correlate with performance ratings even after controlling for stimulus and response length. Thus the validity of the ratings cannot be attributed solely to superficial evaluation of stimulus and response length. Also noteworthy is that normalized visual distance is a strong predictor even after controlling for the number of phonemes correct. The smaller the visual distance between the stimulus and response phonemes, the higher the participant's confidence in correctness of the response.

Discussion

This study showed that participants in both groups, those with normal hearing and those with severe-to-profound hearing losses, were able to judge quite accurately their own speechreading performance. It is likely that the ability to make these judgments is related to normal communication processes that are employed in understanding others and making oneself understood, whether by speechreading or by auditory speech perception. However, the participants with impaired hearing as a group were better able to judge their own performance than were those with normal hearing. At the same time, the range of validity coefficients was extremely wide for both participant groups.

Although partners in conversation doubtless rely to some extent on both knowledge of the language and its constraints and on feedback from their conversation partner to maintain confidence of mutual understanding, the regression analyses in this study suggest that perceivers can also use information in the speech stimulus and their own response to judge their own speechreading accuracy. The regression analyses involving word measures produced R^2 values that ranged between .406 and .526. These analyses suggest that participants were more confident when the stimulus was shorter, and as an independent factor, when their response was longer. Also, the more words actually correct, the more confident the participants. Inasmuch as the participants did not receive feedback telling them when they were correct, the positive relationship between words correct and confidence, independent of number of stimulus words and number of response words, may represent a judgment of the extent to which what they perceived seemed meaningful or plausible to them. This judgment would certainly rely on knowledge of the language in addition to criteria of meaningfulness.

The regression analyses that employed phoneme measures suggest additional possibilities for mechanisms engaged while making confidence judgments. R^2 values for phoneme measures ranged between .425 and .587. The significant β values for stimulus phonemes and response phonemes could represent length judgments that might rely on one or more sources of length information: (a) overall estimates of stimulus and response durations; (b) estimates of the number of syllables visible in terms of mouth opening and closing (Summerfield, 1991) and the number of syllables in the response; and (c) overall estimates of the number of phonemes in the stimulus and response. These are all judgments that could be made independent of knowing what, or if, part of the stimulus was correct.

It is interesting to note that the phonemes correct measure was not consistently significant across participant groups and sentence sets. Phonemes correct is a measure that includes not only the phonemes correct in correct words but any other phonemes that were correct in incorrect words. The inconsistency with which this measure entered the regression analyses may be a result of the fact that the measure incorporates information about correct and incorrect parts of the response to each sentence.

Normalized visual distance is perhaps the most interesting of the phoneme measures entered into the regression analyses. It resulted in consistently high β values across equations, and the zero-order within-subjects correlations involving this factor were quite high, particularly for participants with impaired hearing. Normalized visual distance is the average of perceptual phoneme-to-phoneme distances between stimulus and response across each alignment of an entire stimulus with its associated response. Of particular importance is that correctly perceived phonemes contribute zero distance to the sum of phoneme-to-phoneme distances in the measure. The predictive value of visual distance in the regression analyses suggests that perceivers can judge the overall phonetic similarity of their response relative to the perceived phonetic content of the stimulus. This would seem to be a mysterious ability in the absence of feedback concerning the correct response, as was the case in this experiment. That is, apparently the visual-phonetic distances predict confidence in the absence of absolute knowledge of what the stimulus was. One hypothetical mechanism that could accomplish this would be one that measured the distance between the perceived phonetic input and the word/s in the mental lexicon closest to the phonetic input that is

not completely recognized. This idea is briefly explored below from the perspective of contemporary models of word recognition.

How the Process of Word Recognition Might Contribute to Valid Subjective Performance Ratings

Contemporary information processing models of word recognition (e.g., Luce, 1986; Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Morton, 1969, 1982) incorporate activation processes that might function to measure the distance between the perceived phonetic input and word/s in the mental lexicon closest to the unrecognized phonetic input. In particular, all of these theories hypothesize an early stage of processing in which phonetic stimulation is transformed perceptually into phonetic representations that come in contact with lexical representations stored in long-term memory. Word recognition is a process in which the best match between the incoming phonetic stimulus is isolated from the other stored representations in the lexicon. Representations in the lexicon are said to become activated as a function of their similarity to the incoming phonetic form. Almost all current models of word recognition assume that multiple candidates are activated during word recognition and that activation is graded and not binary (Lively, Pisoni, & Goldinger, 1994). (What mechanism performs selection from among activated candidates is a more contentious issue [see, e.g., Marslen-Wilson & Moss, 1996] and, for purposes here, not the focus of discussion.) In the case of visual speech perception, for which the stimulus is phonetically impoverished, multiple stored lexical forms may become activated, but a unique selection may never be achieved, because no one lexical item may be closest to the stimulus. A plausible hypothesis is that the visual-phonetic distance measure in the current study reflects the ability to employ conscious knowledge of activation levels of words resulting from the stimulus input.

Within much of the word-recognition literature, activation is typically regarded as pre-perceptual and would therefore seem unavailable as the basis for post-perceptual subjective performance ratings. However, there is an area of research that demonstrates the availability of activation to conscious judgments. A well-documented condition is the tip-of-the-tongue (often referred to as "TOT") phenomenon, in which recall of a word (temporarily) fails, but is felt to be imminent (Brown & McNeill, 1966/1970; see A. S. Brown, 1991, for a review). Brown and McNeill conducted an experiment in which definitions for rare words were read and participants were asked to name the defined word. When a tip-of-the-tongue state was induced, the participant was to write down everything about the word that came to mind, including words that were subjectively close to the target word. More words of similar sound than similar meaning were elicited under the TOT condition. Furthermore, there was a rank order correlation of 1.0 between the number of syllables in the similar word and that in the target, and there was an overall high percentage rate of correctly guessing the initial letter of the word. Brown and McNeill attempted to quantify the proximity of similar-sounding words to the target word. They concluded that "a subject at a given distance from the target can accurately judge which of two words that come to mind is more like the target and that he does so in terms of the features of words that appear in generic [TOT] recall" (Brown & McNeill, 1966/1970, p. 291).

Subsequent research on tip-of-the-tongue states supports the idea that the subjective experience is related to lexical activation. Yaniv and Meyer (1987) reported a study in which participants attempted to identify words in the tip-of-the-tongue paradigm. When words could not be retrieved, they rated their "feelings of knowing." Subsequently, the words from the TOT paradigm were presented together with others for lexical decision. The reaction times in the lexical decision task were evaluated in relation to the strength of the feelings of knowing. The results support the notion that unsuccessful attempts to retrieve inaccessible stored information primed the later recognition of the information through a process of spreading activation, and that participants were capable of quite accurately rating the level of activation. Unfortunately, the Yaniv and Meyer experiments did not include collection of response words from participants in the TOT state, so the relationship between strength of knowing and phonological similarity was not assessed.

Meyer and Bock (1992) examined the effects of so-called "interlopers" in the tip-of-the-tongue paradigm. In their experiments, a cue was presented that was similar in sound, in meaning, or not at all. Sound cues were found to be more effective than meaning cues in resolving the TOT state. They interpreted their experiments as support for a partial-activation hypothesis that suggests that "TOT states arise when heightened levels of activation fall short of an adequate level for selection" (p. 723). The utility of the cues similar in sound suggests that activation can spread among words of similar form, enhancing the accurate retrieval of the target.

Of course, the tip-of-the-tongue phenomenon concerns access to the lexicon initially via conceptual or semantic knowledge, not phonetic input. Thus, it remains to be established experimentally that subjective performance ratings in speechreading are based at least in part on conscious access to form-based activation

during spoken word recognition. The work on the tip-of-the-tongue phenomenon supports the notion that activation levels are available at the perceptual or word-recognition end of our speechreading task. It is plausible to us that subjective performance ratings are sensitive to the level of lexical activation between incoming stimulus information and lexical entries that are contacted in memory but not necessarily uniquely selected. Under this hypothesis, speechreaders can experience a sense of being close to recognizing words but failing to achieve confident recognition. This hypothesis leaves open the possibility that activation might also result from semantic relationships between lexical entries. How activation affects speechreading, and whether its effects vary from those during auditory speech perception, is a topic that deserves future investigation.

Conclusion

A substantial majority of individuals with and without hearing impairment are able to provide subjective ratings of their speechreading performance that correlate highly with objective measures. These results are consistent with the literature on estimation of performance in auditory perception and they are interpretable in terms of current psycholinguistic models. This raises the possibility that some aspects of speechreading performance might be studied using subjective ratings rather than actual measures of performance. For example, when group designs are used to evaluate differences among talkers or among experimental conditions, mean subjective ratings might provide efficient and valid measures for making those comparisons. It is unlikely, however, that ratings could be substituted for objective measures in the assessment of individuals. Although the ratings may be highly correlated with objective measures, they are less than perfectly correlated. There would therefore be considerable error if the ratings were used to estimate objective performance. Moreover, there are substantial individual differences in the validity of the ratings. As a consequence, it would not be known, for a given individual, whether the ratings had high, moderate, low, or no validity. For estimation of an individual's performance on a speechreading task, direct methods of measurement are still to be preferred.

Acknowledgments

Portions of this research were presented at the Annual Convention of the American Speech-Language-Hearing Association, San Antonio, TX, November 1992, and at the Summer Institute of the Academy of Rehabilitative Audiology, Salt Lake City, UT, June 1994. The research was supported by NIH grants from the National Institute on Deafness and Other Communication Disorders, DC-00695 and DC-02107.

References

Bernstein, L. E., Demorest, M. E., & Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment. *Journal of the Acoustical Society of America, 95*, 3617–3622.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (1997). *Speech perception without hearing*. Manuscript submitted for publication.

Bernstein, L. E., & Eberhardt, S. P. (1986a). *Johns Hopkins Lipreading Corpus I-II: Disc I* [Videodisc]. Baltimore, MD: Johns Hopkins University.

Bernstein, L. E., & Eberhardt, S. P. (1986b). *Johns Hopkins Lipreading Corpus III-IV: Disc II* [Videodisc]. Baltimore, MD: Johns Hopkins University.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109*, 204–223.

Brown, R., & McNeill, D. (1970). The "tip of the tongue" phenomenon. In R. Brown (Ed.), *Psycholinguistics* (pp. 274–301). New York: The Free Press. (Reprinted from *Journal of Verbal Learning and Verbal Behavior, 5*, 325–337, 1966).

Davis, H., & Silverman, R. (Eds.). (1970). *Hearing and deafness* (3rd ed.). New York: Holt, Rinehart and Winston.

Demorest, M. E., & Bernstein, L. E. (1991). Computational explorations of speechreading. *Journal of the Academy of Rehabilitative Audiology, 24*, 97–111.

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech and Hearing Research, 35*, 876–891.

Demorest, M. E., Bernstein, L. E., & DeHaven, G. P. (1996). Generalizability of speechreading performance on nonsense syllables, words, and sentences: Subjects with normal hearing. *Journal of Speech and Hearing Research, 39*, 697–713.

Demorest, M. E., Bernstein, L. E., & Tucker, P. E. (1997). *Generalizability of speechreading performance on nonsense syllables, words, and sentences: Subjects with impaired hearing*. Unpublished manuscript.

Duffy, S. A., & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech, 35*, 351–389.

Educational Services Department, Digital Equipment Corporation. (1984). *DECtalk DTC01 programmer reference manual*. Maynard, MA: Digital Equipment Corporation.

Finn, J. D. (1974). *A general model for multivariate analysis*. New York: Holt, Rinehart and Winston.

Gray, T., & Speaks, C. (1978). Ability of hearing-impaired listeners to understand connected discourse. *Journal of the American Auditory Society, 3*, 159–166.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Harris, R. J. (1985). *A primer of multivariate statistics* (2nd ed.). Orlando, FL: Academic Press.

Kreul, E. J., Nixon, J. C., Kryter, K. D., Bell, D. W., Lang, J. S., & Schubert, E. D. (1968). A proposed clinical

test of speech discrimination. *Journal of Speech and Hearing Research, 11*, 536–552.

Kruskal, J. B., & Sankoff, D. (1983). An anthology of algorithms and concepts for sequence comparison. In D. Sankoff & J. B. Kruskal (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 265–310). Reading, MA: Addison-Wesley.

Lively, S. E., Pisoni, D. B., & Goldinger, S. D. (1994). Spoken word recognition. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 265–301). New York: Academic Press.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Marslen-Wilson, W., Moss, H. E., & von Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 1376–1392.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25*, 71–102.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29–63.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

Meyer, A. S., & Bock, K. (1992). The tip-of-the-tongue phenomenon: Blocking or partial activation? *Memory & Cognition, 20*, 715–726.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165–178.

Morton, J. (1982). Disintegrating the lexicon: An information processing approach. In J. Mehler, E. Walker, & M. Garrett (Eds.), *On mental representation* (pp. 89–109). Hillsdale, NJ: Erlbaum.

Speaks, C., Parker, B., Harris, C., & Kuhl, P. (1972). Intelligibility of connected discourse. *Journal of Speech and Hearing Research, 15*, 590–602.

Steel, R. G. D., & Torrie, J. H. (1980). *Principles and procedures of statistics: A biometrical approach*. New York: McGraw-Hill.

Summerfield, Q. (1991). Visual perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 117-137). Hillsdale, NJ: Erlbaum.

Van Tasell, D. J., & Hawkins, D. B. (1981). Effects of guessing strategy on speechreading test scores. *American Annals of the Deaf, 126*, 840–844.

Wang, M. D., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*, 663–705.

Yaniv, I., & Meyer, D. E. (1987). Activation and metacognition of inaccessible stored information: Potential bases for incubation effects in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 187–205.

Yanz, J. L. (1984). The application of the theory of signal detection to the assessment of speech perception. *Ear and Hearing, 5*, 64–71.

Yanz, J. L., Carlstrom, J. E., & Thibodeau, L. M. (1985). Self-assessment of communication skills: Toward the development of a new speech audiometric tool. *Ear and Hearing, 6*, 211–215.

Received September 11, 1996

Accepted February 12, 1997

Contact author: Marilyn E. Demorest, Department of Psychology, UMBC, 1000 Hilltop Circle, Baltimore, MD 21250. Email: Demorest@umbc2.umbc.edu